Should Security Researchers Experiment More and Draw More Inferences?

Kevin S. Killourhy Roy A. Maxion

Dependable Systems Laboratory
Computer Science Department
Carnegie Mellon University
5000 Forbes Ave,
Pittsburgh, PA 15213
{ksk,maxion}@cs.cmu.edu

> "Whenever possible, experiments should be comparative. For example, if you are testing a modification, the modified *and* unmodified procedures should be run side by side in the same experiment." *Box, Hunter, and Hunter* [1]

> "It is possible, and indeed it is all too frequent, for an experiment to be so conducted that no valid estimate of error is available. In such a case the experiment cannot be said, strictly, to be capable of proving anything." *R. A. Fisher* [2]

Abstract

Two methodological practices are well established in other scientific disciplines yet remain rare in computer-security research: comparative experiments and statistical inferences. Comparative experiments offer the only way to control factors that might vary from one study to the next. Statistical inferences enable a researcher to draw general conclusions from empirical results.

Despite their widespread use in other sciences, these practices are haphazardly used in security research. Choosing keystroke dynamics as an example to study, we survey the literature. Of 80 papers wherein these practices would be appropriate, only 43 (53.75%) performed comparative experiments, and only 6 (7.5%) drew statistical inferences.

In disciplines such as medicine, comparative experiments and statistical inferences save lives and cut costs. Rigorous methodological standards are required. We see no reason why security research, another discipline where the stakes are critically high, cannot or should not adopt these practices as well. Failure to take a more scientific approach to security research stalls progress and leaves us vulnerable.

1 Introduction

If a science of security were to emerge, we would expect it to resemble other sciences of complex systems: medicine, biology, or even quality control [3, 4]. The data collected and analyzed by security technologies are records of people's behavior, both legitimate and malicious. Packets on a network are largely a manifestation of people, for instance, checking their email, watching movies, or engaging in industrial espionage. A security technology such as a firewall or intrusion detection system (IDS) which analyzes network packets can only be effective if its capabilities allow it to recognize the behavior of users and attackers. It makes sense that a scientific understanding of these security technologies might share features with other sciences that seek to understand complex systems such as people and other organisms.

The research methods for sciences of complex systems all share some similarities. First, they all rely on well-designed comparative experiments. In medicine, a new treatment is compared either to a baseline treatment or to a placebo. Likewise, biology experiments make use of control groups. Comparative experiments allow researchers to control nuisance factors, and to establish causal connections between treatments and outcomes. Second, they all use statistical methods for drawing inferences from experimental results. For instance, statistical hypothesis tests such as the t-test can ascertain whether differences between a control group and a treatment group are significant. Statistical inferences enable a researcher to draw conclusions that hold more generally than a particular experiment. These conclusions, if they withstand scrutiny, form the discipline's body of scientific knowledge.

If we expect that a science of security might be similar to fields such as medicine or biology, then we might justifiably look to those disciplines for good methodological practices. Throughout this paper, we use examples from intrusion detection and keystroke dynamics to demonstrate that nothing special about security research exempts the field from standard scientific practices. In Section 2, we explain the benefits of comparative experiments over one-off evaluations typical of security research. In Section 3, we explain why statistically-based

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2011		

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Should Security Researchers Experiment More and Draw More Inferences?	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,Computer Science Department,5000 Forbes Ave,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Two methodological practices are well established in other scientific disciplines yet remain rare in computersecurity research: comparative experiments and statistical inferences. Comparative experiments offer the only way to control factors that might vary from one study to the next. Statistical inferences enable a researcher to draw general conclusions from empirical results. Despite their widespread use in other sciences, these practices are haphazardly used in security research. Choosing keystroke dynamics as an example to study we survey the literature. Of 80 papers wherein these practices would be appropriate, only 43 (53.75%) performed comparative experiments, and only 6 (7.5%) drew statistical inferences. In disciplines such as medicine, comparative experiments and statistical inferences save lives and cut costs. Rigorous methodological standards are required. We see no reason why security research, another discipline where the stakes are critically high, cannot or should not adopt these practices as well. Failure to take a more scientific approach to security research stalls progress and leaves us vulnerable.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**8**	

Detector	Data Set	Error Rate
A	1	20%
B	2	15%
C	3	10%
D	4	**5%**

(a): Error rates from one-off evaluations

	Data Set			
Detector	1	2	3	4
A	**20%**	**10%**	**0%**	**0%**
B	25%	15%	5%	**0%**
C	30%	20%	10%	**0%**
D	35%	25%	15%	5%

(b): Error rates from comparative experiments

Table 1: Example demonstrating how results from one-off evaluations can be misleading. In Panel (a), where each detector is evaluated on a private data set, Detector D appears best. In Panel (b), where every detector is evaluated on every data set, it is revealed that Detector A is always one of the best and Detector D is never the best.

inferences keep researchers from misinterpreting their own results and each others'. In Section 4, a survey of the literature demonstrates that computer-security research has not yet adopted these practices. In Sections 5–7, we discuss our concern with current practices, review the similar concerns of others', and argue for a more scientific approach to security.

2 Why conduct comparative experiments?

Strictly speaking, what security researchers colloquially call an *experiment* rarely qualifies. Typically, researchers use the term experiment to denote empirical work in contrast to theoretical work. However, in other sciences, not all empirical work qualifies as an experiment:

> **experiment** An investigation in which the investigators have sufficient control of the system under study, in particular to be able to determine the assignment of different units of study to different treatments (conditions or modes of intervention). [5, p.139]

Key to this definition of an experiment is the concept of *comparison*. The purpose of assigning different treatments to different units of study is to compare the effectiveness of those treatments. To avoid confusion with the colloquial meaning of experiment, we use the term *comparative experiment*. For instance, in medicine, a comparative experiment might compare the effectiveness of a new drug (first treatment) to that of the currently recommended drug (second treatment).

We contrast an experiment with what we call a *one-off evaluation*. When proposing a new security technology, researchers often collect a private data set specifically for the purpose of evaluating the technology. They do the evaluation and report the results, thereby demonstrating proof of concept (e.g., that an IDS can detect attacks). Since the evaluation uses a new, never-to-be-reused data set, we refer to it as a one-off evaluation.

2.1 Problem with one-off evaluations

By introducing both a new technology and a new data set, a researcher makes it impossible to separate the effectiveness of the technology with the difficulty of the evaluation. One data set may make an evaluation easier than another. In contrast to a one-off evaluation, one could conduct an experiment by evaluating the new technology against a benchmark data set on which current technologies have already been tested. Then, comparisons between the new and current technologies can be made while keeping the evaluation data constant.

To illustrate the problem with one-off evaluations, consider four researchers who have individually proposed four intrusion detection systems: A, B, C, and D. Table 1 presents the error rates of these four detectors. For the sake of simplicity, assume that every researcher uses the same performance metric (e.g., equal-error rate). In practice, different researchers use different metrics (e.g., misses and false alarms), but that would needlessly complicate the example.

Panel (a) shows the results if all four researchers conduct one-off evaluations. Each researcher collects a private data set and reports evaluation results using that data set. According to these evaluations, Detector A has the highest error, and D has the lowest error of the four. A reader might believe that Detector D is the best detector. However, in the one-off evaluations, each detector is evaluated using a different data set. When multiple factors are allowed to vary in lockstep (e.g., the detector and the data set), statisticians say that they are *confounded* because it is impossible to separate the effects of one from the effects of the other.

Panel (b) shows the results if all four researchers share their detectors and their data, enabling comparative experiments in which each detector is evaluated using each data set. Here, we can see that error rates for all detectors are high when using the first data set and low when using the last data set. In a sense, Data Set 1 is difficult for the detectors while Data Set 4 is easy. The variation in difficulty between data sets overwhelms the differences in detector performance. However, for any given data set (i.e., any column of the table), Detector A has one of the lowest error rates while Detector D has the highest error rate. Detector A appears to be the best detector, not Detector D as the one-off evaluations suggest.

The dangers of one-off evaluations are not confined to contrived examples. For instance, in keystroke dynamics, a detector's accuracy depends substantially on the data used in the evaluation. A neural network's false-alarm rate changed from 1.0% to 85.9% from one evaluation data set to another. On the same two data sets, a k-nearest-neighbor's false-alarm rate changed from 19.5% to 46.8% [6, 7]. The detector and data interaction is so strong that the best-performing detector depends on the data set. This important interaction, which will require more research to understand, cannot be discovered with one-off evaluations.

2.2 Case for comparative experiments

One-off evaluations should be abandoned in almost all circumstances because they are dangerously misleading. Obviously if a new technology purports to solve a totally new problem (e.g., the first IDS ever), there may be no basis for comparison. However, from that point forward, we see no scientific reason not to conduct comparative experiments. Ideally, any new attempt to solve a problem should be compared empirically against prior attempts to solve the same problem.

In practice, comparative experiments of security technologies are hindered by confidential data sets that cannot be shared. While we are sensitive to this issue, it is not a valid excuse. Comparative experiments may be inconvenient, but the dangers of one-off evaluations do not go away simply because they are easier. Suppose, for the sake of discussion, that conferences and journals required comparative experiments. We are confident that many more researchers would find a way to overcome the inconvenience and collect data that they can share.

In the meantime, we should acknowledge that some researchers have recognized the importance of comparative experiments by providing benchmark data sets. The 1998/1999 IDS evaluations constitute comparative experiments [8, 9]. PREDICT and DETER offer data and a testbed, respectively, with which to evaluate technologies under controlled conditions [10, 11]. More recently, benchmarks of other technologies such as worm detection have also been performed [12]. While no benchmark is perfect, we think that research effort is better spent addressing those flaws than continuing to report results of one-off evaluations.

3 Why make statistical inferences?

Regardless of whether a researcher performs a comparative experiment or a one-off evaluation, we believe that he or she has a duty to explain the evaluation results, not just to report them. Suppose a new technology has a 0% error rate in an empirical evaluation. Should the reader infer that the technology will always perform perfectly? A skeptical reader may doubt that any technology works perfectly, but how high could the error rate climb? Should a 1% or even a 10% error rate be anticipated? Obviously, any inferences made from the results of an evaluation will be limited by factors such as the representativeness of the data, but a researcher ought to be able to offer some conclusions within those limitations. If the researcher cannot tell a reader what an evaluation means, why expend the time and cost of conducting what is, in effect, a meaningless exercise?

Detector	Error Rate
A	20%
B	17%
C	15%
D	10%
E	8%
F	6%

Table 2: Error rates for six detectors on an evaluation. Because the table includes no estimates of the uncertainty of the results, they can be interpreted in a variety of ways by different readers (as illustrated in Figure 1).

In other sciences, the process of analyzing empirical results and drawing more general conclusions is known as statistical inference.

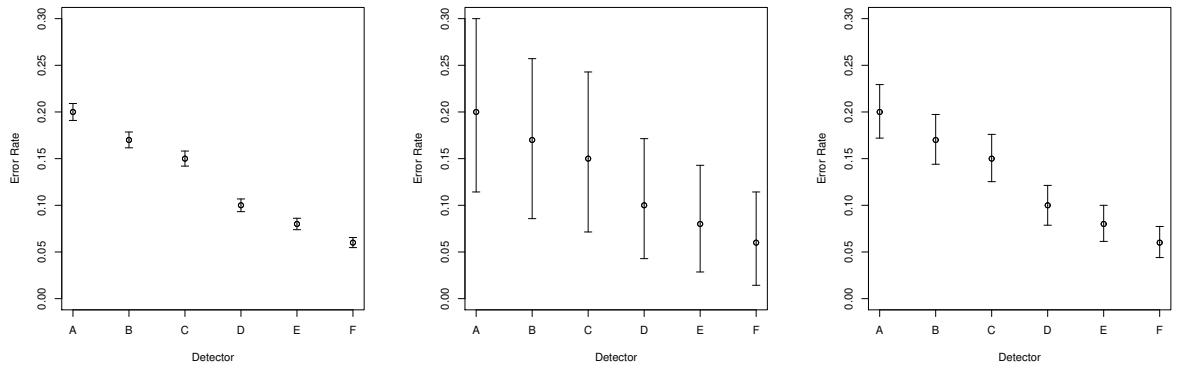
> **inferential statistics** The process of making inferences about a population from findings based on sampled observations. Inferential statistics are used to go beyond the description of the data and to examine hypotheses about underlying research questions. [5, p.199]

For instance, in medicine, a researcher might ask whether a new drug increases the chance of a positive outcome. A drug trial will reveal the percentage of positive outcomes from a sample of patients, but the research question concerns everyone with the disease, not just the sample patients. Inferential statistics offer a means to generalize to the population from the sample.

Computer security has its own pressing research questions. For example, how well will a new intrusion-detection system perform, and is its error rate lower than those of existing IDSes? To answer these questions, a researcher can collect data and conduct an experiment, at which point he knows the error rates of the IDSes on that set of evaluation data. In medicine, inferential statistics would then be used to generalize from the empirical results and answer the research question. However, the security researcher typically reports the empirical results and leaves the research question unanswered.

3.1 Problem with no statistical inference

For the sake of illustration, consider a security researcher who is interested in knowing the relative performance of six detectors. The researcher sets out to



(a) Interpretation 1: Each detector has a different true error rate.

(b) Interpretation 2: All detectors could have the same true error rate.

(c) Interpretation 3: Error rates fall into one of two groups, low or high.

Figure 1: Three different interpretations of the results shown in Table 2. Panel (a) shows one interpretation, where the estimates are all very accurate and so detector F is the best. Panel (b) shows another interpretation, where none of the estimates are very accurate, so no detector is clearly better than the others. Panel (c) shows a third interpretation where A–C are roughly equivalent, D–F are roughly equivalent, but the latter three detectors definitely have lower error than the former. The different interpretations would lead future research in very different directions.

answer the question by conducting an experiment, evaluating each detector using the same data set. Table 2 provides results for this example experiment. The error rates for each of the six detectors are labeled A through F. According to the table, the empirical error rate for detector A was highest, and the error rate for F was lowest. However, there are many different ways to interpret these results.

Figure 1 illustrates three different ways that a reader might interpret the results of the evaluation. One set of readers will see any difference in the empirical error rates as significant. They see the empirical error rates as very accurate estimates of the *true error rates*. In other words, if more data were collected, the evaluation results for each detector would eventually converge to an error rate very close to the empirical one. Panel (a) illustrates the view of readers who believe the estimates are very accurate. The points indicate the empirically calculated error for the six methods, and the small intervals around each point represent where these readers believe the true error rates lie. Since the intervals do not overlap, these people believe all six detectors have significantly different error rates.

Another set of readers looking at the table will think that none of the six error rates are really that different from the rest. Panel (b) illustrates the view of such readers. They see the empirical error rates as very imprecise estimates of the underlying true error. If more data were collected, these people believe that the error rates could change quite a lot. Since these readers believe the intervals are wide and overlap one another, they are unconvinced that the experiment shows any significant difference in the true error rates of the six detectors.

Yet another set of readers might perceive a split between the error rates for the first three detectors and the last three. Based on this perception, they might not see any difference in the true errors for A–C, or for D–F, but they believe there is a significant difference between the A–C group and the D–F group. Panel (c) depicts this interpretation with intervals that overlap within each of the two groups but do not overlap between the groups.

These different interpretations of the results would drive future research in very different directions. If there are significant differences among all six detectors, researchers might focus future attention on further improvements to Detector F, the best detector available. If there are significant differences between groups A–C and D–F, those same researchers should seek to understand whether secondary measures of performance (e.g., speed or resource usage) distinguish detectors D, E, and F. If there are no differences among the detectors, researchers might need to collect more data or ask different research questions.

The danger of a researcher reporting empirical results without drawing statistical inferences is not simply that different readers will have different interpretations of the results. Even if the researcher makes his or her own interpretation clear, some readers will likely disagree. The danger is that the researcher and the readers will not know they disagree. With a clear idea of the researcher's conclusions, future research can undertake to support or disprove those conclusions. When no conclusions are drawn, there is no foundation on which to build future work or to add new knowledge.

3.2 Case for statistical inferences

In our experience, most security researchers are familiar with statistical inference insofar as they know that other sciences use it to distinguish a significant result from one that can be explained by chance. However, it can be challenging to translate hypothesis testing or confidence-interval estimation to security research. The problem arises because security data violate the assumptions behind many statistical inference procedures. Traditionally, the noise is assumed to be Normally distributed and stationary (i.e., not changing over time). In many security domains, these assumptions are probably violated. Security data such as the number of attacks and the frequency of detection are not Normally distributed, and they change over time. To avoid making incorrect conclusions, researchers simply ignore the issue by not making any conclusions at all.

However, computer security is not the first domain to struggle with ill-behaved data. Non-Normal and non-stationary data do exist outside of security data sets, and statisticians have developed statistically sound ways to draw inferences regardless [13]. In many cases, traditional statistical methods have been shown to be surprisingly robust to assumption violations. When necessary, more modern statistical methods like bootstrapping trade computing power for fewer assumptions.

Whatever the difficulty, researchers have a duty to draw inferences and offer their own explanation of their findings, going beyond the mere reporting of empirical results. By not making any inferences or offering any explanations, a researcher has effectively abrogated that duty. An experiment that has not been explained is effectively meaningless, and arguably a waste of time and resources. A preferable alternative would be for researchers to analyze the data and draw conclusions while acknowledging any limitations. Even flawed or limited conclusions provide fodder for future research.

4 Keystroke-dynamics literature review

In the previous sections, we offered arguments in favor of comparative experiments and statistical inferences, and we warned of the dangers of not using them. These arguments will not surprise readers with a statistical background, but our perception is that the security community has either not heard or not heeded these arguments.

Perhaps these practices are not used because security researchers arrive in the field via different paths, including mathematics, formal methods, computer systems, and networks. The diversity of backgrounds has undoubtedly helped the field make advances in many different directions, but it may also have resulted in inadequate exposure to proper experimental research methods.

Whatever the reason, comparative experiments and statistical inferences seem to be the exception rather than the rule in computer security. To establish whether this perception is true, we conducted a survey of a segment of the literature to confirm how often researchers really do conduct comparative experiments and draw statistical inferences.

4.1 Method

We chose to survey papers on keystroke dynamics—the study of whether genuine users and impostors can be distinguished by their typing rhythms—in part because of our familiarity with the topic, but in larger part because some aspects of the topic make it particularly well suited to comparative experiments and statistically based inferences. Keystroke dynamics has been studied for over 30 years [14, 15], and by now there are dozens of existing classifiers against which to compare new proposals. Any reluctance to conduct comparative experiments is difficult to justify. Likewise, standard methods of statistical hypothesis testing (e.g., t-tests or Wilcoxon tests) should be suitable for analyzing evaluation data without much worry about assumption violations. In short, researchers could conduct comparative experiments and make statistical inferences without too much additional investment, and the relative gains to scientific knowledge would be enormous.

To obtain a large and representative sample of keystroke-dynamics research papers, we consulted the *IEEE Xplore* database of articles and conference proceedings published by the IEEE, to which our university maintains a subscription. We conducted two keyword searches for *keystroke dynamics* and *keystroke biometrics*. In total, these two searches returned 101 unique papers: 13 journal articles and 88 conference or workshop papers.

We screened these papers to identify those which described the evaluation of a keystroke-dynamics classifier and reported the evaluation results. This screening excluded 21 papers after which 80 papers remained. The majority of excluded papers were surveys which mentioned keystroke dynamics but did not conduct a technology evaluation. A few of the papers described a new keystroke-dynamics technology but did not evaluate it empirically; consequently, we considered them outside the scope of a review of papers containing empirical evaluations.

We read each of the remaining 80 papers to assess whether a comparative experiment was performed. Specifically, we recognized a paper as having performed a comparative experiment if, in the section describing the evaluation and its results (including tables and figures), the researchers compared the performance of multiple classifiers on the same keystroke-dynamics data set.

	Proportion	95% CI lower	95% CI upper
Comparative experiments:	43 out of 80 = 53.75%	42.2%	65.0%
Statistical inferences:	6 out of 80 = 7.5%	2.8%	15.6%

Table 3: Counts of how many papers perform comparative experiments and draw statistical inferences. In addition to the raw count and the percentage, we include the lower and upper bounds of a 95% confidence interval (assuming a Binomial distribution). The results show that many papers do not perform comparative experiments, and most do not draw statistical inferences.

We consider this definition to be lenient. In fields such as medicine, a new treatment would be compared to an established baseline treatment, not another new treatment. However, we recognized a paper as having a comparative experiment even if two new classifiers were compared. We considered using stricter criteria, but it can be surprisingly tricky to determine whether a classifier is intended to be new (e.g., support vector machines have been independently proposed for keystroke dynamics by several researchers).

While we recognized papers that evaluated multiple classifiers as comparative, we did not recognize papers that evaluated multiple tunings of a single classifier. An exploration of how error rates change with different amounts of training or with different anomaly-score thresholds would not be recognized as a comparative experiment. Otherwise, any paper with an ROC curve would need to be recognized as performing a comparison across different tunings. We felt that including such papers would grossly distort the value of what we aimed to measure.

For the same 80 papers, we also assessed whether a statistical inference was made. Specifically, we recognized a paper as having performed a statistical inference if, in the section describing the evaluation results and analysis (including tables and figures), the researchers reported the results of a hypothesis test (e.g., a p-value) or included confidence intervals. A few authors used the word "significant" without, to the best of our knowledge, meaning it in a statistically precise way. We would not count a statement that a "new classifier was significantly better" as evidence of a statistical inference unless the test procedure was named or a p-value provided. Note that we recognized all statistical inferences, not just those concerning the relative performance of multiple classifiers. For instance, if a researcher performed a hypothesis test to establish that error rates were lower on long passwords than short passwords, we counted it as a statistical inference.

4.2 Findings

A listing of all 101 papers included in our survey is available as an online supplement:
`http://www.cs.cmu.edu/~keystroke/cset-2011`.

Table 3 summarizes the overall results. Of the 80 papers that evaluated keystroke-dynamics classifiers, 43 (53.75%) conducted comparative experiments, and only 6 (7.5%) drew statistical inferences.

The final two columns of the table give 95% confidence intervals. To interpret the confidence intervals, imagine that every paper reporting evaluation results for keystroke-dynamics classifiers were collected and assessed. Based on the whole population, one could calculate the true percentage of papers that include comparative experiments and inferential statistics. Assuming that the 80 papers in our sample are representative (i.e., that papers in *IEEE Xplore* are similar to those published elsewhere), these confidence intervals estimate the regions where those true percentages would lie with 95% probability. The intervals were calculated assuming the counts come from a Binomial distribution with a sample size of 80.

As noted, we believe that our criteria for recognizing a comparative experiment were lenient. Even with that leniency, only a slim majority of papers qualified (53.75%). Many papers simply reported the results of one-off evaluations which are misleading when compared to other results. We are concerned with the state of the field that approximately half of the published results are impossible to compare soundly with other results.

Even more alarming is that only 6 papers (7.5%) draw statistical inferences from their evaluation results. Research that does not draw inferences is difficult to justify in terms of the time and resources expended. Effectively, such research offers results without really explaining what those results mean. That so few papers contribute statistically rigorous conclusions about keystroke dynamics is cause for concern.

5 Discussion

Our intent in surveying the keystroke-dynamics literature is not to criticize individual papers (including some of our own), but to demonstrate the lack of standard scientific practices across the field. The haphazard usage of comparative experiments and the rarity of inferential statistics is alarming and must improve. Our survey was limited to keystroke dynamics, but our experience suggests that one-off evaluations are common and statistical

inferences are rare throughout security research. A failure to adopt these practices wastes resources and stalls progress. In the security arms race, lack of progress decreases security.

In other scientific disciplines, a significant portion of the overall research effort is aimed at improving and refining the scientific methodology used in the field. One famous example is the *Hawthorne effect*, named after a study of worker productivity at Western Electric Company's Hawthorne site. An increase in worker productivity was eventually traced not to the conditions under study but to workers' excitement and enthusiasm over being studied. The discovery spurred changes in the research methods of behavioral science [16]. Some previous studies, which did not account for the Hawthorne effect, were recognized as flawed. Future studies adjusted their methodologies to accommodate the effect.

Computer security should devote a larger portion of its research effort to improving and refining methodology. One-off evaluations without statistical inferences threaten computer-security research just as the Hawthorne effect threatened behavioral-science research. Security results based on one-off evaluations and without proper statistical analysis are dangerously misleading. Just as behavioral research methods changed with the discovery of the Hawthorne effect, we hope that security research methods adopt comparative experiments and inferential statistics going forward.

Of course, even if security researchers begin conducting comparative experiments and drawing statistical inferences, research methods may not be perfect. In this respect, comparative experiments and inferential statistics are necessary but not sufficient. A researcher might conduct a comparative experiment, draw statistical inferences, and still arrive at a fundamentally flawed conclusion. The comparison could be invalid (e.g., due to artifacts in the evaluation data), or the analysis could be wrong. It is easy to see how poor science might happen even with comparative experiments and statistical inferences; it is hard to see how good science will happen with one-off evaluations and no inferential statistics.

A discipline's scientific method arises out of a consensus among researchers. Our hope with this paper is to start a discussion about when comparative experiments and statistical analysis should be used. When is a one-off evaluation okay, and when should a comparative experiment be required? At what point does exploratory work need to become statistically rigorous? We welcome the discussion because, given the current state of security research methods, any discussion is progress.

6 Related work

We are not the first to question the role of experimentation in computer science. Walter Tichy's "Should Computer Scientists Experiment More?" [17] inspired the title of our work. He observed that proper experimentation is lacking throughout computer science research. He identified eight arguments that researchers use to argue against experimentation in computer science. These included the high cost of doing experiments and the sufficiency of mere demonstrations. He also recognized a tendency among computer scientists to award the subject with an ascended status among sciences, effectively rendering the "traditional scientific method" inapplicable. A decade has passed since his paper was written, but we still encounter these sentiments today.

More directly, Peisert and Bishop offered an exemplary scientific method for a science of security [18]. Based on the philosophy of science, they identified three characteristics of a scientific experiment. An experiment must be falsifiable, controlled, and reproducible. These hallmarks of science should certainly be recognizable in any emerging science of security. While we took inspiration from the shared practices of various life sciences rather than philosophy, one could argue that our work and that of Peisert and Bishop are approaching similar positions from different directions. Control and reproducibility are necessary components of a well-designed comparative experiment, and falsifiability is integral to statistical inferences made via hypothesis testing (e.g., establishing criteria by which the null hypothesis is rejected).

A small portion of computer-security research has already taken aim at improving methodology. For instance, in computer security, Sommer and Paxson have offered guidance on evaluation methodology when machine-learning algorithms are involved [19]. More generally, Kurkowski et al. established best practices in the evaluation of mobile ad-hoc network protocols [20]. Their methodological recommendations include establishing standard simulation environments for ease of comparison and the estimation of confidence intervals. They conducted a survey of papers similar to ours and identified many shortcomings, establishing that computer science, not just computer security, faces a methodological challenge.

7 Summary

Comparative experiments are necessary to control lurking variables and rule out alternative explanations, yet in a survey of 80 security papers only 43 (53.75%) actually used a comparative experiment. The rest performed one-off evaluations that have been shown to produce incomparable and misleading results. Admittedly, comparative experiments in computer security are made difficult by a lack of shared data, but that inconvenience does not make the alternative (one-off experiments) a better option.

Inferential statistics are widely used in other sciences for drawing conclusions and explaining the results of an experiment, but in computer security, they are almost never used. In the survey of 80 security papers, only 6 (7.5%) drew statistical inferences. Researchers have a duty to explain what their experimental results mean, otherwise, the experiment had no point. Assumption violations and other complexities can make statistical inference a challenge but not an insurmountable one.

Sciences such as medicine demand comparative experiments and statistical inferences. The stakes in computer security are just as high, but scientific methods are seen as a bonus rather than a requirement. With this paper, we hope to promote discussion of these practices, what problems arise when they are not used, and when they should be required. Our opinion is that comparative experiments and statistical inferences are necessary for a science of security.

Acknowledgments

The authors are indebted to David Banks who reviewed our statistical arguments. We are also grateful to the anonymous reviewers for their comments and suggestions.

This work was supported by National Science Foundation grant number CNS-0716677 and by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office.

References

[1] G. E. P. Box, J. S. Hunter, and W. G. Hunter, *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley, New York, 2nd ed., 2005.

[2] R. A. Fisher, *The Design of Experiments*. Hafner Publishing, NY, 1966.

[3] D. Mayer, *Essential Evidence-Based Medicine*. Cambridge University, UK, 2nd ed., 2010.

[4] H. M. Wadsworth Jr., K. S. Stephens, and A. B. Godfrey, *Modern Methods for Quality Control and Improvement*. John Wiley & Sons, NY, 1986.

[5] Y. Dodge, ed., *Oxford Dictionary of Statistical Terms*. Oxford University, New York, NY, 2003.

[6] S. Cho, C. Han, D. H. Han, and H.-I. Kim, "Web-based keystroke dynamics identity verification using neural network," *Journal of Organizational Computing and Electronic Commerce*, vol. 10, no. 4, pp. 295–307, 2000.

[7] K. S. Killourhy and R. A. Maxion, "Comparing anomaly-detection algorithms for keystroke dynamics," in *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2009)*, (June 29–July 2, 2009, Estoril, Lisbon, Portugal), pp. 125–134, IEEE Computer Society, Los Alamitos, CA, 2009.

[8] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Webber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, and M. A. Zissman, "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation," in *DARPA Information Survivability Conference and Exposition (DISCEX 2000)*, vol. 2, (January 25–27, 2000, Hilton Head, SC), pp. 12–26, IEEE Computer Society, Los Alamitos, CA, 2000.

[9] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "Analysis and results of the 1999 DARPA off-line intrusion detection evaluation," in *Recent Advances in Intrusion Detection (RAID 2000)*, (October 2–4, 2000, Toulouse, France), pp. 162–182, Springer-Verlag, Berlin, 2000.

[10] Predict: Protected Repository for the Defense of Infrastructure Against Cyber Threats, 2010. `https://www.predict.org/`.

[11] DETERlab Testbed: cyber-DEfense Technology Experimental Research laboratory Testbed, 2010. `http://www.isi.edu/deter/`.

[12] S. Stafford and J. Li, "Behavior-based worm detectors compared," in *Recent Advances in Intrusion Detection (RAID 2010)*, (September 15–17, 2010, Ottawa, Ontario, Canada), pp. 38–57, Springer-Verlag, Berlin, 2010.

[13] A. Madansky, *Prescriptions for Working Statisticians (Springer Texts in Statistics)*. Springer-Verlag, NY, 1988.

[14] G. Forsen, M. Nelson, and J. Raymond Staron, "Personal attributes authentication techniques," Tech. Rep. RADC-TR-77-333, Rome Air Development Center, October 1977.

[15] R. S. Gaines, W. Lisowski, S. J. Press, and N. Shapiro, "Authentication by keystroke timing: Some preliminary results," Tech. Rep. R-2526-NSF, RAND Corporation, May 1980.

[16] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston, 2002.

[17] W. F. Tichy, "Should computer scientists experiment more?," *Computer*, vol. 31, pp. 32–40, May 1998.

[18] S. Peisert and M. Bishop, "How to design computer security experiments," in *5th World Conference on Information Security Education (WISE 2007)*, (June 19–21, 2007, West Point, NY), pp. 141–148, Springer, New York, 2007.

[19] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *2010 IEEE Symposium on Security and Privacy*, (May 16–19, 2010, Berkeley, CA), pp. 305–316, IEEE Computer Society, Los Alamitos, CA, 2010.

[20] S. Kurkowski, T. Camp, and M. Colagrosso, "MANET simulation study: The incredibles," *Mobile Computing and Communications Review—Special Issue on Medium Access and Call Admission Control Algorithms for Next Generation Wireless Networks*, vol. 9, pp. 50–61, October 2005.